

A Critical Examination of Test-Scoring Methods

BF
21
A7
no.80

Anderson

Class 151.2

Book No. 40932



Northeastern University
Library

A CRITICAL EXAMINATION OF TEST-SCORING METHODS

BY
ROSE G. ANDERSON, Ph. D.

ARCHIVES OF PSYCHOLOGY
EDITED BY R. S. WOODWORTH

No. 80

NEW YORK
AUGUST, 1925

ARCHIVES OF PSYCHOLOGY

COL. UNIV. P. O., NEW YORK CITY. The Subscription price is five dollars per volume of about 500 pages. Volume I comprises Nos. 1-10, Volume II, Nos. 11-18, Volume III, Nos. 19-25, Volume IV, Nos. 26-32, Volume V, Nos. 33-39, Volume VI, Nos. 40-46, Vol. VII, Nos. 47-52, Vol. VIII, Nos. 53-58, Vol. IX, Nos. 59-63, Vol. X, Nos. 64-68, Vol. XI, 69-73, Vol. XII, 74-78, Vol. XIII, 79. The available numbers are as follows:

2. On the Functions of the Cerebrum: The Frontal Lobes: SHEPHERD IVORY FRANZ. 50c.
3. Empirical Studies in the Theory of Measurements: EDWARD L. THORNDIKE. 50c.
4. Rhythm as a Distinguishing Characteristic of Prose Style: ABRAHAM LIPSKY. 50c.
5. The Field of Distinct Vision: W. C. RUDINGER. 70c.
6. The Influence of Bodily Position on Mental Activities: ELMER E. JONES. 50c.
7. A Statistical Study of Literary Merit: FREDERIC LYMAN WELLS. 30c.
8. The Relation Between the Magnitude of the Stimulus and the Time of Reaction: SVEN FROBERG. 35c.
9. The Perceptual Factors in Reading: FRANCIS MARION HAMILTON. 50c.
10. Time in English Verse Rhythm: WARNER BROWN. 70c.
11. The Hearing of Primitive Peoples: FRANK G. BRUNER. \$1.00.
12. Studies in Development and Learning: EDWIN A. KIRKPATRICK. \$1.00.
13. The Inaccuracy of Movement: H. L. HOLLINGWORTH. 80c.
14. A Quantitative Study of Rhythm: HERBERT WOODROW. 60c.
16. On Certain Electrical Processes in the Human Body and their Relation to Emotional Reactions: FREDERIC LYMAN WELLS and ALEXANDER FORBES. 40c.
17. The Relative Merit of Advertisements: EDWARD K. STROOG, JR. \$1.00.
18. Attention and Movement in Reaction Time: J. V. BRETHERS. 50c. (Cloth, 75c.)
19. An Empirical Study of Certain Tests for Individual Differences: MARY THEODORA WHITLEY. \$1.25. (Cloth, \$1.50.)
20. Visual Acuity with Lights of Different Colors and Intensities: DAVID EDGAR RICE. 50c. (Cloth, 75c.)
21. The Curve of Forgetting: C. H. BEAN. 45c. (Cloth, 70c.)
23. Reaction Time to Retinal Stimulation A. T. POFFENBERGER, JR. 70c. (Cloth 95c.)
24. Interference and Adaptability: ARTHUR JEROME CULLER. 75c. (Cloth, \$1.00.)
25. Reaction to Multiple Stimuli: JOHN WELHOFF TODD. 60c. (Cloth, 85c.)
26. A Study in Incidental Memory: GARRY C. MYERS. \$1.00 (Cloth, \$1.25.)
27. A Statistical Study of Eminent Women: COLEA SUTTON CASTLE. 80c. (Cloth, \$1.03.)
28. The Mental Capacity of the American Negro: MARION J. MAYO. 60c. (Cloth 85c.)
29. Experimental Studies in Judgment: H. L. HOLLINGWORTH. \$1.25. (Cloth \$1.50.)
30. The Psychological Researches of James McKeen Cattell: A Review by Some of His Pupils. \$1.00. (Cloth, \$1.25.)
31. Fatigue and Its Effects upon Control: ISAAC EMERY AUST. 60c. (Cloth, 85c.)
32. The Transfer Effects of Practice in Cancellation Tests: MELVIN ALBERT MARTIN. 60c. (Cloth, 85c.)
33. The Intellectual Status of Children Who are Public Charges: J. L. STENQUIST, E. L. THORNDIKE, and M. R. TRARUE. 50c. (Cloth, 75c.)
34. The Relation of Quickness of Learning to Retentiveness: DARWIN OLIVER LYON. 50c. (Cloth, 75c.)
35. The Overcoming of Distraction and Other Resistances: JOHN J. B. MORGAN. 75c. (Cloth, \$1.00.)
36. The Psychology of the Negro—An Experimental Study: GEORGE O. FERGUSON, JR. \$1.25. (Cloth, \$1.50.)
37. The Effect of Distraction on Reaction Time: JOHN E. EVANS. \$1.00. (Cloth, \$1.25.)
38. The Effect of Humility on Nervousness and on General Efficiency: LORELE IMA STECHER. 90c. (Cloth, \$1.15.)
39. The Mechanism of Controlled Association: MARK A. MAY. 75c. (Cloth, \$1.00.)
40. Recitation as a Factor in Memorizing: ARTHUR I. GATES. \$1.00. (Cloth, \$1.25.)
41. Mental Fatigue during Continuous Exercise of a Single Function: THOMAS RUSSELL GARTH. 85c. (Cloth, \$1.10.)
42. A Psychological Study of Trade-Mark Infringement: RICHARD H. PAYNTER, JR. 85c. (Cloth, \$1.10.)
43. Individual Differences and Family Resemblances in Animal Behavior: HALSEY J. BAGG. 70c. (Cloth, \$1.00.)
44. Experimental Studies in Recall and Recognition. EDITH MULHALL ACHILLES. 90c. (Cloth, \$1.25.)
45. The Morphologic Aspect of Intelligence: SANTE NACCARATI. 70c.
46. Psychological Examinations of College Students: F. EDITH CAROTHERS. \$1.25.
47. The Effects of Practice on Judgments of Absolute Pitch: EVELYN GOUGH. \$1.25.
48. An Experimental Study of Silent Thinking: RUTH S. CLARK. \$1.40.
49. Some Empirical Tests in Vocational Selection: HERBERT W. ROGERS. 75 cents.
50. Adenoids and Diseased Tonsils: Their Effect on General Intelligence. MARGARET CONN. \$1.00
51. An experimental Study of the Factors and Types of Voluntary Choice: ALFRED H. MARTIN. \$1.50.
52. Some Well-Known Mental Tests Evaluated and Compared: DOROTHY RUTH MONTGOMERY. 80 cents.
53. Mood in Relation to Performance: ELIZABETH T. SULLIVAN. \$1.00
54. The Influence of Incentive and Punishment upon Reaction Time: ALBERT M. JOHANSON. 80c.
55. Psychological Tests Applied to Factory Workers: EMILY THORP BURN. \$1.25.
56. A Study of the Relation of Accuracy to Speed: HENRY E. GARRETT. \$1.25.
57. An Experimental Study of Hunger in Its Relation to Activity: TOMI WADA. \$1.50.
58. Individual Differences as Affected by Practice: GEORGINA STICKLAND GATES. \$1.00.
59. Studies in Industrial Psychology: ELSIE OSCHIN BREGMAN. 90 cents.
60. The Mental Status of Psychoneurotics: ALEXANDER D. TENDLER. \$1.25.
61. Effects of Attention on the Intensity of Cutaneous Pressure and on Visual Brightness: SIDNEY M. NEWHALL. \$1.25.
62. The Measurement of Motor Ability: EVELYN GARFIELD. 90 cents.
63. Race Differences in Inhibition: ALBERT L. CRANE. \$1.50.
64. Individual Differences in Incidental Memory: SADIE MYERS SHELLOW. \$1.25.
65. Character Traits as Factors in Intelligence Test Performance: WILLIAM M. BROWN. \$1.25.
66. A Study of the Sexual Interests of Young Women: F. I. DAVENPORT. \$1.25.
67. The Psychology of Confidence: WM. CLARK TROW. \$1.25.
68. Experimental Studies of College Teaching: HAROLD E. JONES. \$1.25.
69. The Influence of Treatment for Intestinal Toxemia on Mental and Motor Efficiency: ALICE E. PAULSEN. \$1.00.
70. A Study of Suggestibility of Children: MARGARET OTIS. \$1.50.
71. The Value of Praise and Reproof as Incentives for Children: ELIZABETH B. HULLOCK. \$1.00.
72. Attention and Interest in Advertising: HOWARD K. NIXON. \$1.25.
73. An Experimental Study of Thinking: EDNA HEIMREDER. \$1.75.
74. Estimation of Time: ROBERT AXEL. \$1.00.
75. Physical Motor and Sensory Traits: R. C. SOMMERVILLE. \$1.00.
76. Measurement of Emotional Reactions: DAVID WECHSLER. \$1.75.
77. Tested Mentality as Related to Success in Skilled Trade Training: THEODORA M. ARTEL. \$1.25.
78. Aggressive Behavior in a Small Social Group: E. M. RIDDLE. \$1.75.
79. The Memory Value of Advertisements: EDITH R. BRANDT. \$1.25.
80. A Critical Examination of Test-Scoring Methods: ROSE G. ANDERSON. \$1.00.

A CRITICAL EXAMINATION OF TEST-SCORING METHODS

BY
ROSE G. ANDERSON, Ph. D.

ARCHIVES OF PSYCHOLOGY
EDITED BY R. S. WOODWORTH

No. 80

NEW YORK
AUGUST, 1925

e

Psych
BF
21
A7
no. 80

ACKNOWLEDGMENTS

The writer wishes to take this opportunity to express her appreciation to Dr. F. Kuhlmann for making accessible the large mass of data which has accumulated during the course of the development and standardization of these tests and for suggestions in regard to this study. She is also indebted to Dr. A. T. Poffenberger for helpful conferences during the progress of this investigation. Thanks are also due Miss Helen Platt for assistance of a statistical nature.

CONTENTS

CHAPTER		PAGE
I	Introduction	5
II	General Plan of the Study	9
	Tests and Subjects Used	9
	Methods of Scoring	12
III	Presentation of Results	18
	Comparisons of Methods with Thirteen-Year Group	18
	Comparisons of Methods for Thirteen- and Four- teen-Year Groups	22
	Comparisons with Eight-Year Children	25
	Conclusions for Investigation to This Point	27
IV	Multiple Correlation Weighting	29
V	Summary	40
	Bibliography	42
	Appendix	43

6 Feb. 1932

40932

A Critical Examination of Test-Scoring Methods

I

INTRODUCTION

Developments in the actual construction and the practical use of intelligence tests, especially group intelligence tests, have unquestionably far exceeded those in the theory as to the principles to be considered in their construction and in the interpretation of results obtained by them. This is generally recognized as an undesirable state of affairs. Many, however, resign themselves to it as the inevitable consequence in a pioneer field. Many others are too much preoccupied by the practical demands made upon them to be able to devote any study to the theoretical aspects. Still others deprecate this tendency to the extent that they belittle the results actually accomplished. The following criticism by Ruml¹ illustrates this attitude. This writer maintains that "relative to the time and number of people devoted to work with mental tests, the results have been astonishingly meagre in theoretical value. Extensive collection of data through mental tests began without the necessary antecedant and contemporaneous development of point of view, hammering out of contradictions in concepts and hypotheses, and elimination of ambiguities in common everyday words and ideas. It is probable that many of the failures of mental tests can be traced to our present inadequate theoretical foundations." Although this statement underestimates the value and importance of results obtained and overlooks the growing body of contributions of theoretical importance, it lays a needed stress on the desirability for closer attention to and consideration of the theoretical aspects of the construction and use of tests.

The particular aspect of mental tests which is investigated in this study is the matter of weighting, *i.e.*, the determination of whether it is necessary or desirable for the best results to

¹ Ruml, Beardsley, "The Need for an Examination of Certain Hypotheses in Mental Tests," *Jr. Phil. Psych. & Sc. Meth.*, 17: 57-61, 1920.

weight the number of items or elements passed in the separate tests of a group-intelligence scale, or whether the simple sum of the number of items passed on all the tests is as likely to give results which are as reliable and valid. To state the problem concisely: Can greater reliability or validity or both be secured by weighted scores than by raw scores?

Theoretically it has been urged that tests should be weighted and a number of methods of weighting have been proposed. Thorndike, as chief of the statistical unit in the initial experiment with the group examination in the army, reported that tests should be weighted according to their correlations with officers' ratings of the mental ability of their men (*i.e.*, according to their correlations with the criterion) and according to their intercorrelations.² McCall and others have recommended weighting tests according to the variability of their scores.³ Henmon has stated⁴ that "the assignment of equal weight to each test and to each item in the individual test is surely wrong." Woodrow has recommended weighting tests according to their discriminative capacity, *i.e.*, the extent to which they distinguish between unselected children of successive age-groups.⁵ Kohs⁶ states that "for testing purposes the unrefined measure obtained for a test should be first transferred into P. E. or S. D. values. These only can give us the absolute values necessary in scaled data." A number of others have recommended this method of expressing scores on tests. Will has proposed the use of a "kental" principle similar to percentiles, but taking into consideration the amount of difference between scores as well as the relative position.⁷ Thurstone has derived formulae from which it is possible to ascertain how much to penalize for errors in a single test in order to obtain the highest correlation with a criterion.⁸ He demonstrates that in some tests partial credit should be allowed for errors instead

² *Memoirs*, National Academy of Sciences, Vol. XV, 1921, p. 316.

³ McCall, Wm. A., "*How to Measure in Education*," MacMillan, 1922, p. 30.

⁴ Henmon, V. A. C., "Intelligence and Its Measurement: A Symposium," *Jr. Ed. Ps.* 12: 195-198, 1921.

⁵ Woodrow, Herbert, and Grace Arthur, "An Absolute Intelligence Scale: A Study in Method," *Jr. App. Ps.* 3: 118-137, 1919.

⁶ Kohs, Samuel C., "Percentile Norms for Scaling Data," *Jr. Ed. Ps.* 9: 101-102, 1918.

⁷ Will, Harry S., "A Method of Commensurating Mental Measurements," *Jr. Ed. Res.* 5: 139-153, 1922.

⁸ Thurstone, L. L., "A Scoring Method for Mental Tests," *Psych. Bull.*, 16: 235-240, 1919.

of penalizing for them and concludes that "the correlations between tests and the criterion are seriously affected by assuming a scoring method at random."

Thus it is seen that there is general agreement as to the value of some method of weighting and many suggestions as to which method is preferable or advisable. Few experimental studies have been made in which the actual comparison of different methods with each other and with raw scoring have been made. The following excerpt from a report by Kelley⁹ is pertinent to this matter: "Any worker presenting a new procedure should definitely recognize that it is impossible to prove the superiority of his method by reporting data interpreted by means of his method alone. Superiority is a relative matter and it is necessary to compare a method with alternative methods before its superiority can be established."

A good deal of consideration was given to the matter of weighting by the psychologists responsible for the army mental tests.¹⁰ Reference was made above to Thorndike's recommendation in regard to this matter. Fairly early in the initial experiment with the tests it was decided that any convenient method of weighting gave a prophecy not much inferior to that obtained by the best possible weighting. In a check of some of the early results it was found that the system of weighting in use yielded correlations with officers' ratings which were not appreciably better than those secured by raw scores with these ratings. For a group of 900 unselected men the correlation between raw scores and weighted scores was .994. In the case of a more homogeneous group of 300 men this correlation was .93.¹¹ It was then decided to weight tests according to the variability of their scores, the measure of variability used being the interquartile range. The correlation of these weighted scores with the raw scores for 2,856 unselected men was .993. Because of this close agreement and the large amount of time required for weighting scores and the added possibilities of errors in scoring, raw scores were thereafter substituted for weighted scores.¹²

⁹ Kelley, Truman L., "Report of the Sub-committee on Statistical Methods of the Standardization Committee," *Jr. Ed. Res.* 4: 77-78, 1921.

¹⁰ *Op. cit.*

¹¹ *Op. cit.*, p. 340.

¹² *Op. cit.*, p. 342.

A study was made by West¹³ of "The Significance of Weighted Scores" in the course of the construction of a test of ability to reproduce thought in passages read. The aim of this study was to determine the advantage, if any, of weighting the separate items in a single test. West's results are from 45 high school students from grades 8 to 12. The per cent of pupils failing each item was translated into "percentile" values. These values were then used as one set of scores and were correlated with the simple point scoring. The conclusion reached was that the chances were greatly in favor of the simple point scoring being as correct as the weighted score. It was thought that the type of test might have a bearing on the conclusion. The same comparison was then made with army alpha and an analogy test. The close agreement of the scores by the different methods led to the final conclusion that "weighting will serve its best purpose in assisting in a scaled arrangement of material even though the items scored be not scored by different values."

Thus, although there is general agreement among psychologists in favor of weighting, experimental evidence at hand seems to be in favor of raw scores being practically as valuable as weighted scores. In order to determine whether this conclusion had been reached because of the particular method of weighting used and the particular comparisons made, and whether it would hold for any of several possible methods, it was thought worth while to make an exhaustive comparison of several different methods on the same tests with groups large enough for the results to be significant of actual differences.

¹³ West, Paul W., "The Significance of Weighted Scores," *Jr. Ed. Ps.*, 15: 302-308, 1924.

II

GENERAL PLAN OF THE STUDY.

Tests and Subjects Used.

Before taking up the matter of what tests were used, a few explanatory remarks will be advisable in regard to a group intelligence scale constructed by the writer under the general direction of Dr. F. Kuhlmann of the Research Bureau of the Minnesota State Board of Control. One feature of this scale which distinguishes it from group intelligence scales on the market is the provision for a separate group of tests for each chronological age from six to fourteen years. There is some overlapping of tests from age to age, but in general the tests in each booklet or scale are tests which fit the children of that age better than children of any other age. The original plan was to place the tests in an age-group where about two-thirds of the cases pass about half the trials. Although it was not possible to adhere to this plan as closely as was first intended, the general advantage of the plan has been attained. This advantage is that every test then yields a measure of the ability in that task for the great majority of cases. In scales which use the same tests over a range of several grades some of the tests are too difficult for the youngest children and others are too easy for the oldest children, with the consequent result that zero and perfect scores are obtained for these tests for these cases, neither of which is satisfactory as a measure of ability. The plan in giving the group tests in the schools is to obtain the average chronological age of the children in that grade and to use the booklet which corresponds closest to this average age unless there is evidence that the average mental ability of the group is decidedly above or below the average chronological age. In order to provide for the range of ability which is met with in any one grade, each booklet has been standardized on a wide range of ages, *e.g.*, mental ages from 4 to 9 years may be obtained with the six-year booklet, from 4 to 13 years with the seven-year booklet, from 4 to 13 years with the eight-year booklet, etc.

That a scale arranged in this way would be preferable to one of the usual type in which the same tests are used over a range of several grades has been recognized by others. Buckingham¹ calls attention to the fact that children and young people at different levels of development differ in kind of intelligence as well as in degree. He concludes therefore that a single mental test cannot be successfully used over a wide range of intellectual levels. He adds that "as a product of future investigation we hope for a hierarchy of tests each appropriate to a given level and each linked with the other by ascertained relationships."

Mohlman² in a study of the discriminative value of the sub-tests of a group intelligence scale concluded that her results with the Pressey Scale add to the existing evidence that tests showing a high degree of value for testing the intelligence of immature subjects of varying mental ability and of low scholastic standing, are likely to be tests which possess little or no value for measuring adults of high scholastic standing and of superior mental ability.

Vincent³ in a study of intelligence test elements found that types of elements which discriminate between the intelligent and unintelligent at one level of difficulty do not always do so at another and observes that it is left to the theory of scale construction to apply this fact.

During the standardization of this scale it was necessary to give all tests to a very wide range of grades. It was therefore possible to select three towns in which the children of any one age although scattered through a wide range of grades had taken all of the tests included in the final booklet for that age. Thus the thirteen-year group, although distributed from grades III to IX inclusive with a few in special classes, had all taken the tests in the present thirteen-year booklet with slight exceptions. The five third grade cases had not taken the last three, and the thirteen fourth grade cases had not taken the last two tests in the booklet. A satisfactory method was worked out for estimating scores on these tests. In every case the tests were given through the entire school system, so that

¹ Buckingham, B. R., "Intelligence and Its Measurement: A Symposium," *Jr. Ed. Ps.*: 12: 271-295, 1921.

² Mohlman, Dora Keene, "The Discriminative Value of the Sub-tests of a Group Intelligence Scale," *Sch. & Soc.* 14: 399-400, 1922.

³ Vincent, Leona, *A Study of Intelligence Test Elements*. Teachers College, Columbia University, 1924.

as far as it is possible to obtain unselected age-groups, this was done.

The major part of this study is based upon the results of unselected thirteen-year olds for the tests in the thirteen-year booklet. There are 17 tests in this booklet. In order to have an even number, one was dropped which was similar to another in the booklet. The number of items in the separate tests varies from 10 to 30. Samples of the tests used are supplied in the appendix. A number are of the type used in other scales, although the form in which they are used is often different, and a number are original tests not used elsewhere. In order to make comparisons between two successive age-groups the fourteen-year group was scored on these same tests using the same scoring as for age thirteen.

Since the tests used for the lower ages are tests containing few items, and since there was a larger number of zero and perfect scores for these ages, it was thought that conclusions reached for the older age-groups would not necessarily hold for younger age-groups. An additional group of unselected 8-year olds was consequently added to the groups under consideration, the tests used being those in the eight-year booklet. The three groups used were as follows: 382 thirteen-year olds, 361 fourteen-year olds, and 393 eight-year olds. These are, with a few exceptions, all the children of these ages in three school systems having a total attendance of 4617. The exceptions are three thirteen-year olds who were absent for part of the tests and three eight-year olds. Two of the latter were in the fourth grade and one in the fifth and had therefore not taken all the tests in the eight-year booklet.

Since it will be of interest later, the distributions by grades and the average ages and S.D.'s of the groups used are given in Tables I and II.

TABLE I
DISTRIBUTION BY GRADES OF AGE-GROUPS USED

	Sp.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Year 8 (7-6 to 8-5)		97	251	45								
Year 13 (12-6 to 13-5)	13			5	16	35	103	146	54	10		
Year 14 (13-6 to 14-5)	10		1		7	11	38	98	134	53	8	1

TABLE II
AVERAGE AGES AND S.D.'S OF AGE-GROUPS

	8 yr.	13 yr.	14 yr.
Av.	7 yr. 11.8 mo.	13 yr. 0 mo.	14 yr. 0 mo.
S.D.	3.35 mo.	3.40 mo.	3.36 mo.

Methods of Scoring

The problem of scoring methods first arose for the writer during the construction of the group-intelligence scale described above. In considering the question of what method of scoring to use, it was decided that a preliminary study of four methods would determine the method to be adopted. Dr. Kuhlmann made a comparison of the four following methods: (1) raw scoring, one point for each item passed in a test; (2) point scoring, *i.e.*, giving the same maximum of fifteen points to each test, the credit for a single item being in inverse proportion to the number of items in the test; (3) sigma scoring, the per cent failing each number of items from 1 to n in each test was translated into the corresponding S.D. score, the appropriate S.D. score then being substituted for the number of items passed on a given test; (4) sigma $\times r_c$ scoring, the above sigma scores were weighted by the correlation of the sigma scores for each test with the composite score on all the tests. The last method might be objected to on the grounds that a spurious correlation is obtained due to the fact that the test being correlated with the composite is in each case included in the composite. This would be an important objection in case only a few tests comprised the composite, but since the number of tests in the composite was 17, it is not a serious one. The result of this preliminary study was the adoption of the fourth of these methods since I.Q.'s obtained by it correlated slightly higher with individual test I.Q.'s than those for any of the other methods. The Kuhlmann Handbook of Mental Tests¹ was used for the individual examinations.

In the present study five methods of scoring were used for the 13- and 14-year groups. These methods will be referred to throughout this account by the following numbers and names: (1) raw scoring, (2) sigma scoring, (3) sigma $\times r_c$ weighting ($\sigma \times r_c$), (4) sigma $\times r_I$ weighting ($\sigma \times r_I$), and (5) multiple correlation weighting. The first method requires no explana-

¹ Warwick & York, 1922.

tion since it simply uses the sum of the items passed on all the tests as the final score. Sigma scoring is the same method referred to above. An example will suffice to illustrate this method and the two following methods. Table III, column 2, gives for test 15 the per cents of unselected thirteen-year olds failing to pass the number of items indicated. Table VI in the

TABLE III
SCORES FOR TEST 15 FOR DIFFERENT METHODS

No. of Items	Per cent Failing	Sigma Value	Sigma X 10	σ_{Xr_c} $r_c = .63$	σ_{Xr_I} $r_I = .54$
1.	5.0	1.37	14	9	7
2.	6.8	1.52	15	9	8
3.	11.0	1.79	18	11	10
4.	15.1	1.97	20	13	11
5.	22.6	2.25	23	14	13
6.	34.3	2.60	26	16	14
7.	46.5	2.92	29	18	16
8.	61.8	3.30	33	21	18
9.	74.8	3.66	37	23	20
10.	87.5	4.14	41	26	22
11.	93.5	4.51	45	28	24
12.	96.1	4.75	48	30	26
13.	99.1	5.31	53	33	29
14.	99.7	5.63	56	35	30
15.	100.	6.00	60	38	32

appendix of Rugg's *Statistical Methods Applied to Education*² was used to translate these per cents into sigma values. This table was used because the zero point has been transferred from the mean to minus 3 S.D., thus doing away with negative values. In order to have numbers more easily handled, these sigma values were multiplied by ten and fractions eliminated. These values are shown in column four of Table III. For method 2, if a child passed a certain number of items in a test, he was given the corresponding sigma value as his score. For example, five items passed in this test would be given a score of 23.

It will be seen that this method allows a child of a certain age credit for any number of items passed according to the difficulty, as determined by the per cent of his age-group which is able to pass that number of items. It does not, however, take into consideration the value of the test in contributing to the result sought, namely, a measure of mental ability. It will readily be conceded that, theoretically, better results

² Houghton Mifflin Co., 1917.

should be obtained if tests are weighted according to the extent to which they do this. The third method or $\sigma \times r_c$ scoring is an attempt to do this. It was known that the total test score on the tests had a high validity as determined by comparisons of group test results by this scale with individual test results for a large number of cases. It was therefore decided to weight each test by the correlation of the test with the composite. It was not possible to take the inter-correlations between tests into consideration in this weighting because they were not then available. Table III, column five, gives the values for each number of items passed after the sigma values have been multiplied by this correlation. The correlations of the separate tests with the composite are given in Table IV. It should be mentioned that Pearson's product-moment method has been used for all correlations in this study. Attention should also be called to the fact that all these correlations are on single age-groups and that they would be correspondingly higher if they were for groups including a wider range of cases.

TABLE IV

CORRELATION OF THE SEPARATE TESTS WITH THE COMPOSITE SCORE (r_c)
AND GRADE POSITION (r_I)

Tests	1	2	3	4	5	6	7	8
r_c	.47	.59	.53	.70	.66	.57	.56	.62
r_I	.54	.60	.52	.67	.56	.60	.56	.67
Tests	9	10	11	12	13	14	15	16
r_c	.65	.45	.56	.53	.37	.79	.63	.60
r_I	.70	.38	.49	.61	.40	.58	.54	.56

The fourth method or $\sigma \times r_I$ was an effort to accomplish the same thing arrived at by method 3, but to improve upon this method by using a criterion outside of the tests. The problem then arose of finding an adequate criterion. Individual test results were available for only a small per cent of these cases and these were highly selected. Other group test results were not available. Grade position was decided upon as the best available criterion. With children of different ages and in individual cases, this could not be considered a satisfactory criterion. But in a group of unselected children of this age, the majority have had equal or approximately equal educa-

tional opportunity and the most important single factor which is responsible for grade position is general mental ability. This conclusion is corroborated by the army test psychologists, who conclude that "the high correlations of the army *a* test with grade position of unselected age groups offer striking proof of the validity of this test as a measure of educability."³ As pointed out by them, the correlations would probably be higher but for certain constant errors in school grading, principally the tendency to over-promote dull children and under-promote bright children. The correlations of these separate

TABLE V
SCORING CHART FOR SIGMA SCORING METHOD⁴

	Tests																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
Items																		
1	13	9	14	12	11	12	13	9	14	16	16	9	11	19	14	19		
2	13	14	15	13	12	14	14	10	15	19	19	10	12	23	15	22		
3	14	15	17	14	13	17	15	12	16	22	22	11	13	27	18	25		
4	14	16	19	16	15	20	15	14	17	25	25	12	15	30	20	27		
5	14	18	21	17	17	23	16	16	18	28	28	14	18	33	23	31		
6	16	21	24	18	20	28	17	19	20	31	31	15	22	37	26	34		
7	19	23	28	19	22	34	17	23	22	33	34	17	26	41	29	37		
8	21	27	33	20	26	43	18	26	26	36	37	19	28	45	33	40		
9	26	30	40	22	28	45	19	29	30	40	41	21	32	49	37	42		
10	31	37	46	24	32	52	20	34	33	45	49	23	35	56	41	44		
11	36			27	36		21	40	37			24	38		45	46		
12	43			31	41		22	43	42			26	41		48	49		
13				34	46		23	60	47			27	46		53	50		
14				39	52		24	60	53			28	50		56	53		
15				45	57		25	60	60			30	53		60	56		
16							26					31						
17							26					32						
18							27					33						
19							28					35						
20							29					36						
21							30					37						
22							31					38						
23							32					39						
24							33					40						
25							34					41						
26							35					43						
27							36					46						
28							37											
29							38											
30							39											

³ *Op. cit.*, p. 335.

⁴ No increase in score from one item to another does not indicate that there was no increase in the per cent failing the next higher number of items, but that this increase was not sufficient to raise the sigma value by one point.

tests with this criterion vary from .38 to .70 with an average of .56. The time limits of these tests are two or three minutes. In view of this fact and the limited range, these correlations are very satisfactory. In the last column in Table III the scores are given for this test by this method. Table IV contains the correlations of the separate tests with grade position. To illustrate the way the scores run for the various tests, the scores are given in Table V for the sigma scoring method. In scoring a booklet, the number of items passed on any test was first secured. A glance at the tables then gave the sigma score for this number of items which was then entered on the score-sheet. Tables were arranged in the same form for both the other methods, the scores in each case being these sigma scores multiplied by the appropriate correlation.

The last method using weights derived by multiple correlation will be described and considered by itself in a later section.

Other methods were considered and abandoned for various reasons. One of these is weighting according to the variabilities of the scores, *e.g.*, using multipliers as weights which would equate the standard deviations for the different tests. This was not used because the S.D.'s for all the tests but two were so nearly alike and because the sigma scoring method used accomplishes much the same result. Weighting according to the discriminative capacity of the tests was not used because it was decided to use this method as a check on the validity of the methods used instead of as a means of deriving weights. Thurstone's method would not be applicable because results were in the form of number of trials correct on each test with no record of errors.

In order to avoid complicating the comparisons to be made, these were not made between mental ages secured by the different methods, but between total scores with their variabilities always taken into consideration. The matter of the form in which to express final ratings is a problem in itself and is left to future consideration.

Although in the actual use of this scale, children in the third grade would be rated according to their score in the tests given to the third grade, and those in the ninth grade by the tests given to the ninth grade, this plan has not been followed in this study. All children of the same age have been rated on

the same tests irrespective of the grade they were in. There is, therefore, no reason to suppose that results of this study will not be applicable to other group test scales and also to other types of scales.

III

PRESENTATION OF RESULTS

Comparisons of Methods with Thirteen-Year Group.

Several different comparisons were made in order to obtain a measure of the comparative validity of scores by the different methods. The extent to which tests distinguish between groups known to be different is an indication of their validity. It will be granted that thirteen-year children in special classes for subnormal children and in grades III to V are inferior in general mental ability to thirteen-year children in the VIIIth and IXth grades. If, in the comparison of these two groups, one method distinguishes between them more successfully, it justifies the conclusion that this method has greater validity. This comparison was made for the four methods for these two groups. The average score for each group for each method was found. The difference between the averages and the P.E.'s of these differences were then calculated. The ratio of the difference to the P.E. of the difference was finally used in the comparison of the different methods. Table VI includes the results for this comparison.¹

TABLE VI
COMPARISON OF RETARDED AND ACCELERATED PUPILS

Method	Av. Score Grades (VIII-IX) (Sp., III-V)		Diff.	P.E. of Diff.	Ratio of Diff. to P.E. of Diff.
Raw	185.31±2.34	86.74± 5.08	98.57	3.77	26.13
σ	592.58±7.76	316.49±14.55	276.09	11.12	24.82
$\sigma \times r_c$	352.28±4.64	187.27± 8.92	165.01	6.78	24.34
$\sigma \times r_I$	320.31±4.28	168.30± 7.69	152.01	5.97	25.45

The results of Table VI show that one method distinguishes about as successfully as another between these two groups

¹ All calculations in this study have been made by the use of tables and the Monroe and Burroughs calculating machines and have been carefully checked. For the sake of consistency all calculations have been carried to three decimal places even where this degree of accuracy was not considered necessary or significant. In general, results are reported to two decimal places.

with any evidence in favor of any one method in favor of raw scoring.

These two groups comprise the lowest 18% and the highest 17% of all the cases as determined by their position in a grade noticeably above or below that expected for this age. These surveys were made at the end of the academic year so that position in grades VI and VII would be considered normal for ages twelve years, six months to thirteen years, five months. Incidentally, attention may be called to the success with which these tests, by any method of scoring, distinguish between these groups. These ratios may be compared with the corresponding ratio of 22.12 obtained by the army tests for the difference between 300 English privates and 703 officers.²

But differences due to refinements in scoring methods might not be expected to manifest themselves in groups as widely divergent as these. This same comparison was therefore made between grades (Sp., III-V) and grades (VI-IX) which would presumably not differ so much, since the average of the accelerated group would be lowered by the addition of grades VI and VII. This comparison is given in Table VII. The groups are again very successfully distinguished but there is no evidence that one method is superior to another in making this distinction.

TABLE VII

COMPARISON OF RETARDED WITH AVERAGE AND ACCELERATED PUPILS

Method	Grades (VI-IX)	Av. Score (Sp., III-V)	Diff.	P.E. of Diff.	Ratio of Diff. to P.E. of Diff.
Raw	154.49±1.66	86.74± 5.08	67.75	3.61	18.78
σ	502.92±4.80	316.49±14.55	186.43	10.33	18.04
$\sigma \times r_c$	296.45±2.90	187.27± 8.92	109.18	6.32	17.26
$\sigma \times r_I$	270.88±2.65	168.30± 7.69	102.58	5.48	18.71

Still another comparison was made using the differences between the thirteen-year-olds in grades V, VI, VII and VIII. The other grades were not used because of the small numbers of cases in them. These differences would again be smaller than those in the last comparison and would be expected to show slighter changes from method to method. The ratio of

² *Op. cit.*, p. 328.

the difference to the P.E. of the difference between the averages for each two successive grades is given in Table VIII.

TABLE VIII
RATIO OF DIFFERENCE TO P.E. OF DIFFERENCE FOR GRADES (V-VIII)

Grades	VI-V	VII-VI	VIII-VII
Method			
Raw	5.30	9.79	16.61
σ	6.67	7.88	12.45
$\sigma \times r_c$	5.52	8.53	12.76
$\sigma \times r_I$	5.56	8.87	12.36

Reliable differences are still found between these groups for all of the methods. As was expected, more pronounced differences are found between the different methods especially for the three upper grades, with greater evidence than before that the last three methods have no advantage over the raw-score method, since for these three grades the differences are less reliable for these methods than for the latter.

Another measure which is often used as a criterion of validity is the amount of overlapping between groups when one is known to be superior to the other. This is usually expressed in the terms of the per cent of the lower group which reaches or exceeds the median of the higher. This comparison was made for the two groups made up of grades (Sp., III to V) and (VI to IX). There are slight differences in the amounts of overlapping. The following numbers of the lower group reached or exceeded the median of the higher group: raw score, 1; sigma scoring, 4; sigma $\times r_c$ scoring, 3; and sigma $\times r_I$ scoring, 3. In per cents these numbers are respectively: 1.4, 5.8, 4.3, and 4.3. The results in per cents for this same comparison between the grades (V to VIII) are given in Table IX.

TABLE IX
PER CENT OF LOWER GRADE REACHING OR EXCEEDING THE MEDIAN OF THE NEXT HIGHER

Grade	V	VI	VII
Method			
Raw	25.7	27.2	11.0
σ	29.0	24.3	12.3
$\sigma \times r_c$	25.7	24.3	11.0
$\sigma \times r_I$	22.9	24.3	13.0

The apparently larger differences in the fifth grade are due to the smaller number of cases in this group. There is a difference of only one more case for sigma scoring and one less case for sigma x r_I scoring than for the other two methods.

As a measure of the discriminative capacity of a test between groups, Woodrow³ has suggested using the differences between the averages of the two groups in terms of the average standard deviation. His formula is:

$$D.V. = \frac{Av_2 - Av_1}{\frac{1}{2} (\sigma_1 + \sigma_2)}$$

These discriminative values were obtained for the different methods for each of the groups considered above. Although they throw no further light on the comparisons made they are included (Table X) in case they may be more meaningful to some readers than some of the results already given.

TABLE X
DISCRIMINATIVE VALUES FOR DIFFERENT SCORING METHODS

Groups	(Sp., III-V) vs. (VIII-IX)	(Sp., III-V) vs. (VI-IX)	V vs. VI	VI vs. VII	VII vs. VIII
Method					
Raw	3.24	1.89	.70	.87	1.69
σ	3.02	1.81	.86	.68	1.31
$\sigma \times r_c$	2.97	1.74	.73	.74	1.35
$\sigma \times r_I$	3.08	1.86	.84	.79	1.34

Still another method of arriving at the comparative validity of these methods was to determine the correlation of the scores with grade position. These correlations are as follows:

Raw	Scoring.....	.80
σ	"	.79
$\sigma \times r_c$	"	.79
$\sigma \times r_I$	"	.79

Even in the fourth method in which the tests have been weighted according to their correlations with grade position, the correlation of the final score with grade position has not been raised. The correlation of the scores for each of the other methods with raw scores was also obtained. In all three cases this correlation was .97. Table XI is included to show how closely the scores by two of the methods agree.

³ *Op. cit.*

TABLE XI
RELATION BETWEEN RAW SCORES (Y) AND SIGMA SCORES (X) FOR 382
UNSELECTED 13-YEAR-OLDS

	0-49	50	100	150	200	250	300	350	400	450	500	550	600	650	700
210														3	4
195												1	13	9	
180											1	26	11		
165											28	19			
150										21	41	3			
135									13	44	3				
120								2	39	9					
105							1	23	11	1					
90						1	7	9							
75						3	7								
60						6	5								
45					1	4									
30				2	1										
15		2	2	1											
0-14	2	2	1												

This concludes the comparisons made within the thirteen-year group to determine which method has greater validity. Any indication of superiority revealed to this point has been in favor of the raw score method. But any differences found have not been large enough nor consistent enough to warrant a final conclusion.

Comparisons of Methods for Thirteen- and Fourteen-Year Groups

Probably one of the most frequently used means of judging tests has been the extent to which they distinguish between successive age-groups. It is maintained by some, however, that an increase in score by an older group may be so largely due to longer exposure to environmental conditions and advantages in training as to invalidate this method of comparison. It is admitted that this factor makes it difficult to gauge

how much of an increase in score is due to growth of intelligence, independent of training, and that any comparison which fails to take this factor into consideration is open to criticism. It will be acceded, however, that the difference revealed by tests of this nature would be due at least in part to growth of intelligence. Moreover, for the matter under consideration the more important thing is not the interpretation of any difference found, but the determination of whether this difference is more evident by one method of scoring than another. The comparison of methods was consequently made for the entire thirteen- and fourteen-year groups. The results of this comparison are given in Table XII. The ratios of the difference between the averages for the two groups to the P.E. of this difference, the discriminative values determined in the same manner as above, and the per cents of the thirteen-year group reaching or exceeding the median of the fourteen-year group are given for each of the methods.

TABLE XII

COMPARISONS OF DIFFERENT METHODS FOR UNSELECTED 13- AND 14-YEAR AGE GROUPS¹

Method	Ratio of Diff. to P.E. of Diff.	Discrim- inative Value	Per cent of Overlapping
Raw	2.80	.138	42.9
σ	3.76	.186	40.1
$\sigma \times r_c$	3.56	.179	40.3
$\sigma \times r_I$	8.22	.407	27.7

For the first time these comparisons consistently show an improvement of the other three methods over the raw score. While the ratios of 3.76 and 3.56 are hardly reliable differences, they are noticeably nearer to what is considered a reliable difference (a ratio of 4) than the ratio of 2.8 for the raw scores. And for the last method the ratio of 8.2 leaves no question as to the advantage of this method over any of the others for distinguishing between these two groups. This conclusion

¹ Attention should be called to the fact that these differences are not measures of the validity of this scale. For these comparisons, the 14-year-olds are scored on the 13-year tests with the sigma values based on 13-year-olds. In actual practice the 14-year-olds would be scored on 14-year tests with sigma values based on 14-year-olds. Resulting comparisons for the two groups might then be very different.

is further brought out by the relations between the discriminative values and the amounts of overlapping. A change in the latter from 43% to 28%, due only to a change in the method of scoring, is very significant. This part of the study then leads to the conclusion that for purposes of distinguishing between these successive age-groups, any of the last three methods is superior to the raw score method, with the last method considerably superior to either of the other two.

The Relative Reliability of the Different Methods

The final comparisons made for these four methods with these two groups was made for the purpose of ascertaining whether there was any evidence that greater reliability could be expected from one method than from another. In order to have a measure of reliability for the different methods of scoring, the group of sixteen tests was divided into two half-scales by arbitrarily putting alternate tests in one half-scale and the remaining ones in the other half-scale. Thus one half-scale consisted of the even numbered tests and the other of the odd numbered tests. The scores on the half-scales were then correlated with each other and a coefficient of reliability obtained for the entire scale by means of the Brown-Spearman Formula. Kelley gives this formula in the following form⁵ when the reliability of the scale is determined from the scores on the two halves:

$$r_{11} = \frac{2r}{1 + r} \frac{\frac{1}{2} \frac{I}{II}}{\frac{1}{2} \frac{I}{II}}$$

This method of obtaining a reliability coefficient is frequently used when only a single form of a test is available and when the test has not been repeated. It is admittedly somewhat of a make-shift for getting a true reliability coefficient, since for this formula to be strictly applicable each separate item in each test should be paired with a comparable item in the test with which it is correlated. This is obviously not the case when the two half-scales are built up in the above manner. Kelley states⁶ that in most mental and educational test work the true reliability coefficient will be less than that obtained

⁵ Kelley, Truman Lee, *Statistical Method*, MacMillan, 1923, p. 206.

⁶ *Op. cit.*, p. 203.

for repeated tests and greater than that obtained for comparable tests. Presumably the reliability coefficient obtained in the manner described above would also be lower than the true reliability coefficient. Since the information desired here is the relative reliability *i.e.*, whether a higher reliability may be expected for one method than for another, this method may be considered satisfactory for the purpose for which it is used. Table XIII contains the reliability coefficients for the two age-groups.

TABLE XIII
RELIABILITY COEFFICIENTS FOR DIFFERENT SCORING METHODS

	13-Yr.-Group	14-Yr.-Group
Method		
Raw	.93	.94
σ	.92	.93
$\sigma \times r_c$	.93	.93
$\sigma \times r_I$	.93	.94

In view of these coefficients, it cannot be concluded that any of these methods would be likely to have a higher reliability than any other.

Comparisons with Eight-Year Children

As mentioned above, it was considered advisable to extend this investigation to include a younger age group. Because of the difficulty of getting sustained attention on the part of younger children over longer periods of time, it has been found that the best type of test for the younger ages is one with only a few items. The majority of tests for the younger ages are tests of three to six items. Because of this fact a larger per cent of perfect scores are obtained for these ages, since a narrower range of ability is provided for. There is also a large per cent of zero scores. These are due partly to the greater likelihood of younger children not comprehending the directions for tests because of fluctuations in attention or the unusualness of the situation. Zero and perfect scores are also in part due to the greater disparity in ability from age to age of children of the younger ages and therefore the greater frequency in the same group of children so widely separated in ability that some fail to comprehend the task while others get a perfect score.

In view of the above considerations, conclusions drawn from the investigation with the older age-groups might not hold for the younger ages. The eight-year group was therefore included in order to make the study more comprehensive. The tests used are those in the eight-year booklet of the scale described above. Most of these tests are in the form of pictures or forms, the response in most cases consisting of making a dot, line, or other appropriate mark. Only two of the tests require a response in the form of writing words or letters. One of these and two others require the reading of words to perform the task and in these cases the words are read to the group in the course of giving the instructions. There are fifteen tests in this booklet. Since an even number was again desired, one test for which some scores were missing was dropped from the group used for this study.

The methods used for this group were: (1) raw scoring, (2) sigma scoring and (3) sigma times r_c scoring. It was not possible to use the fourth method or sigma times r_1 scoring because a criterion was lacking. Grade position could not be used for a criterion for this group because of the narrow range of distribution and because position in a certain grade for this age is much less likely to be determined by ability than it is for an older group which has been in school longer.

Of the 393 cases, 25 per cent were in the first grade, 64 per cent in the second grade, and 12 per cent in the third grade. There is no question that a larger per cent of the first grade group are mentally dull than in the case of either of the other groups and that a larger per cent of the third grade group are

TABLE XIV
COMPARISONS OF METHODS FOR EIGHT-YEAR-GROUPS—GRADES I AND III

	Ratio of Diff. to P.E. of Diff.	Discrimi- native Value	Per cent of Overlapping
Method			
Raw	21.04	2.42	2.0
σ	20.21	2.23	2.0
$\sigma \times r_c$	22.15	2.51	2.0

mentally superior than in the case of either the first or second grade groups, and that the difference between these two groups is again indicative of the difference in their mental ability. Consequently the difference between these two groups

was again used in comparing the three methods. The same comparisons were made as in the case of the older groups. Table XIV contains the results of the three comparisons. The measure of overlapping is again the per cent of the lower group which reaches or exceeds the median of the higher.

These groups are again practically as successfully distinguished by one method as by another, with very slight indications in favor of the superiority of the $\sigma \times r_c$ scoring.

The reliability coefficients were secured in the same manner described above. These coefficients for the different methods were:

Raw	scoring.....	83
σ	"	84
$\sigma \times r_c$	"	85

There is then slightly more evidence in favor of the $\sigma \times r_c$ weighting for this group than in the case of the older groups.

Conclusions for Investigation to This Point

Any conclusions made with regard to the value of any method of scoring must take into consideration the expenditure of time for both the standardization of tests and for scoring tests which that method requires in comparison with others. When this factor is constant, any slight demonstrable superiority of one method over another would warrant its adoption in preference to the other. But where this slight superiority entails the expenditure of much greater time and effort it is highly questionable whether the adoption of that method should be recommended because of this superiority.

In this study all comparisons within the thirteen-year group failed to disclose any superiority of any of the other methods over the raw score method. When differences were found, they were usually slight and in favor of the raw score method. In the study with the eight-year group, there was a little indication that the $\sigma \times r_c$ scoring was superior to the raw score method, but the differences found were so slight that it would be difficult to determine whether they were due to a real superiority of this method or to chance factors not taken into consideration. Even if they could be shown to be due to the former, they are scarcely large enough to warrant the extra time and labor necessary to obtain them.

The general conclusion from the results for these two groups would be that the results by the raw score method have as high validity and reliability as the results by any of the three methods of weighting used and that in view of this and the greater ease of obtaining norms and of scoring, this method is preferable to any of the others.

The results for the comparison between the thirteen- and fourteen-year groups, however, would lead to the conclusion that any of the three methods of weighting used are superior to the raw score method, with the fourth or sigma $\times$ r₁ method so much in the lead that its adoption might be advisable.

The final decision will rest on whether the difference between retarded and advanced children within an unselected age-group has greater validity as a criterion of mental ability than the difference between successive age-groups. It is the view of the writer that differences between these two former groups with the age-factor constant are more significant of differences in mental ability. For conclusive evidence that one method is preferable to another, its superiority should be demonstrated in the comparisons made between these groups as well as between the successive age-groups. This should be the case even if it is not granted that differences within the age-group are more significant. It would appear, therefore, that although one or the other of these methods of weighting, principally the fourth or sigma $\times$ r₁ weighting, may be slightly superior to the raw score method, the evidence that this is the case has not been sufficient to warrant the recommendation of the adoption of any of these methods in preference to the raw score method. This is especially true when the extra expenditure of time required for the other methods is taken into consideration.

The reader should not infer from this conclusion, that improvements cannot be made over the raw score method nor that efforts should be discontinued to seek these improvements. The implication is rather that efforts toward improvement could probably be made more profitably if made in some other direction than by means of the methods considered above.

IV

MULTIPLE CORRELATION WEIGHTING

The derivation of weights by the technique of multiple correlation is admittedly the "best possible weighting." But because of the laboriousness of this method, it is more often recommended than used. It was thought that it would be of value in this study to determine how great the advantages of this method are over the simpler methods of weighting and over the raw score method.

Instead of the usual method of obtaining partial correlations and building up a regression equation with the desired number of variables, there are two alternative methods which give the desired weights and multiple correlation with the expenditure of appreciably less time and labor, especially in the case of a larger number of variables. One of these is Kelley's method of successive approximation.¹ By means of this method it is possible to estimate weights for tests according to their correlations with the criterion, and by using these weights to derive another set of weights and at the same time to determine the multiple correlation coefficient for the weights used. These derived weights are then used and the process is repeated until it results in new weights which are identical with those used in obtaining them. When this occurs, it is proof that the regression coefficients have been found.

The other method is one which has been devised by Toops. The correlation coefficient obtained by this method is not a true multiple correlation coefficient, but a very close approximation to it. Since this method is as yet unpublished, Dr. Toops' explanation of it has been reprinted from another study which used this method.²

"The problem of securing the *maximum* predictive value from a *minimum* number of tests resolves itself into the problem of determining that test, U, of a number of available tests, which will yield by the technique below presented a maximum

¹ Kelley, Truman Lee, *Statistical Method*. MacMillan, 1923, p. 302.

² Garfiel, Evelyn, *The Measurement of Motor Ability*, Archives of Psychology, 1923.

multiple ratio correlation coefficient when combined at the proper weight with an already existing weighted test composite, C, which already has a maximum correlation with the criterion. This involves the determination of the correlation of test U with the composite, C; the determination of the weight β , of the test U, such that when the deviations in terms of standard deviations in Test U are weighted by that amount (β), the multiple ratio correlation coefficient r_{Ic}' (I being the criterion) of the new test composite at that point shall be a maximum by this method of computation.

"At the outset, that test of the n available tests which correlates highest with the criterion is taken as the "backbone" test, and is named Test C (in this series: Test 9); the correlation coefficient r_{Ic} (in this series .70) is a maximum at this point of building up a scale. If the gross scores in Test C are now given a weight of $\frac{1.000}{\sigma_c}$, there is for any test, U, a weight ($\beta_{u/c}$) such that when the gross scores in Test U are multiplied by $\beta_{u/c}/\sigma_u$ the multiple correlation coefficient r_{Ic}' of the two-test scale correlated with the criterion shall be maximum. But some of the tests, when weighted at their own individual $\beta_{u/c}$ weights, will yield higher multiple correlation coefficients than others. Hence, the ordinary formula for multiple correlation is solved for each of the remaining ($n-1$) tests in turn:

$$r_{Ic}' = \sqrt{\frac{r_{Ic}^2 + r_{Iu}^2 - 2r_{Ic} \cdot r_{Iu} \cdot r_{uc}}{1 - r_{uc}^2}} \quad (1)$$

That test which yields the maximum multiple correlation coefficient is now called Test U. The weight of Test U with respect to Test C weighted 1.000 is

$$\beta_{u/c} = \frac{r_{Iu} - r_{Ic} \cdot r_{uc}}{r_{Ic} - r_{Iu} \cdot r_{uc}} \quad (2)$$

We now consider that theoretically the gross scores on Test U have been added at the proper weight to the gross scores of Test C; our problem is then to find by formula the weight of a new Test, U', such that when its gross scores are added to the now existing test composite, C' (the gross scores of which are considered as now being weighted 1.000), the multiple ratio correlation coefficient shall be a maximum for all the possible remaining ($n-2$) tests. It is not necessary actually to

combine the gross scores and compute the necessary correlation coefficients $r_{u'c'}$ since a formula obtains the same result:

$$r_{u'c'} = \frac{\sum r_{u'y} \cdot W_y}{\sqrt{\sum W_y^2 + 2\sum r_{xy} \cdot W_x \cdot W_y}} \quad (3)$$

in which, $r_{u'y}$. W_y is the sum of the single-products of all the correlations of Test U' that occur in a column U', each respectively multiplied by the weight of the Test W_y of the row for all the rows of the test composite C' which enter into a double symmetrical intercorrelation table, which is being built up as the test composite is being built up. This test composite at this point consists of Tests C and U, whence the two correlation-products are $r_{u'c}$ (1.000) and $r_{u'u}\beta_{u/c}$.

" $\sum W_y^2$ is the sum of the squares of the weights in the test composite at this time: namely, here $(1.000)^2 + (\beta_{u/c})^2$.

" $\sum r_{xy} W_x W_y$ is the sum of all the double-products when each of the intercorrelations of tests in the test composite at this time are respectively multiplied by the weight of the row and column in which they are found in an intercorrelation table, namely, here, r_{uc} (1.000) $\cdot \beta_{u/c}$.

"With this equation solved for all the remaining (n-2) variables, resort to formula (1) determines which test will yield the maximum multiple ratio correlation coefficient, $r_{Ic''}$. This test, when determined, is called U". The weight of U" in the multiple regression equation is given by the formula:—

$$\beta' = \sqrt{\sum W_y^2 + 2\sum r_{xy} \cdot W_x \cdot W_y} \times \frac{r_{Iu'} - r_{Ic'} \cdot r_{u'c'}}{r_{Ic'} - r_{Iu'} \cdot r_{u'c'}} \quad (4)$$

"The quantity under the radical does not enter into formula (2) for the reason that the standard deviation of the original test C is 1.000, when measured in terms of its own standard deviation. When adding on Test U' the composite C' has a standard deviation of its own which must be considered, and which this radical expression takes full account of. Equation (4) is the perfectly general expression of the weight at which a new test U' is added to an already existing test composite C'. By repetitions of the procedure involved in adding Test U' as above outlined, one may determine in succession the fourth, fifth, sixth tests, and so on. The multiple ratio correlation coefficient at each point is an index of the efficiency of the scale. Soon a point of diminishing returns is reached, where the addition of a test adds but little to the value of the multiple

ratio correlation coefficient at that point, and the value available will approach the value which we would receive from the inclusion of the entire n tests. At this point the test can be considered complete.

"Any two or more of the tests can be used for the scale by cutting off from the composite any number of tests which are added last.

"The multiple ratio correlation coefficient is not a true multiple correlation coefficient but is a very close approximation to it."

This latter method was decided upon for the derivation of the desired weights because of the possibility of seeing just what the effect of adding each test to the composite would be, because it permitted the number of tests in the team to be limited to any number thought desirable without any extra labor, and because it was possible to build up from the tests used, two comparable half-scales to use in getting the reliability coefficient.

The first pre-requisite for the use of this method is the correlations of the tests with the criterion and all inter-correlations between tests. For sixteen tests the required number of inter-correlations is 120.³ These inter-correlations for the 382 thirteen-year-olds together with the correlations with the criterion and the S.D.'s for each test are included in Table XV. The inter-correlations were furnished to four decimal places. The correlations with the criterion have also been carried to this number of decimal places, since this degree of accuracy is required in the computation of the weights and the multiple correlation. Grade position was again used as the criterion.

The customary procedure in using this method is to continue to add tests to the composite until the addition of further tests does not increase the correlation of the composite with the criterion. At this point the team of tests is considered complete. Beginning with test 9 as the "back-bone" test, the other tests were added in the order in which they are entered in Table XVI. This table contains the weights (β'), the gross score weights (β' divided by the S.D. for the test), the final weights used, the cumulative correlation after the addition of

³ The writer was fortunate in being able to have these correlations computed by Dr. Hull of the University of Wisconsin on his recently invented computing machine. Dr. Hull writes of these correlations that he doubts if a large scale computation of correlation coefficients has ever been as precisely done as these.

TABLE XV
CORRELATIONS WITH CRITERION (1), INTER-CORRELATIONS AND S.D.'s.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	S.D.'s.
1	.5408															2.5601
2	.6014	.4063														2.5035
3	.5205	.3989	.4882													2.5596
4	.6722	.4120	.6477	.5457												3.5544
5	.5609	.4782	.5893	.5842	.6541											2.5870
6	.6037	.4527	.5096	.5690	.5783											1.9824
7	.5636	.4058	.5146	.4745	.5537	.4917	.5319									8.6005
8	.6682	.5100	.5355	.4732	.5687	.5408	.4705	.5015								2.5678
9	.6992	.4818	.5986	.5590	.7061	.6171	.5535	.5031	.5641							3.0807
10	.3846	.3193	.3542	.2870	.4150	.4644	.4108	.3961	.4149	.3894						3.0259
11	.4950	.3378	.3804	.2972	.3906	.3786	.3935	.4150	.4642	.4311	.3070					2.8516
12	.6107	.4880	.3871	.3464	.4211	.4165	.4056	.4605	.5037	.5022	.2368	.4387				6.6473
13	.4037	.3792	.4040	.3419	.3767	.3562	.3685	.3491	.4333	.4356	.2743	.3319	.3785			3.0146
14	.5793	.3704	.4944	.4427	.5890	.5686	.4171	.5311	.6026	.5851	.4980	.4644	.4373	.4520		2.3628
15	.5414	.4293	.4317	.5565	.5585	.4956	.4820	.4496	.5545	.5545	.3365	.3951	.4059	.3634	.4774	2.8557
16	.5628	.4202	.3505	.4501	.3970	.4508	.5093	.4652	.4712	.2928	.4490	.5915	.3764	.5202	.4093	3.2610

each test, and the cumulative time. The derived weights are divided by the S.D.'s since the scores must be used in terms of their variability. The gross score weights obtained thus are too cumbersome to use in calculation. The final weights were obtained from these by first multiplying by 100 and then dividing through by 4. This reduces the weights to a manageable size, but retains approximately the same relations between them.

TABLE XVI
WEIGHTS AND CORRELATIONS FOR WEIGHTED TEST COMPOSITES

Test	Weight β' Sigma	Gross Score Weights	Final Weights	Cumulative Correlation	Cumulative Time
9	$\frac{1.000}{3.0807}$	.3246	8	.6992	2 min.
8	$\frac{.8495}{2.5678}$	.3308	8	.7738	5
12	$\frac{.6561}{6.6473}$	.0987	2	.8012	7
6	$\frac{.6246}{1.9824}$	.3151	8	.8182	9
4	$\frac{.5783}{3.5544}$	.1627	4	.8257	11
16	$\frac{.3628}{3.2610}$	.1113	3	.8291	14
2	$\frac{.3289}{2.5035}$	.1314	3	.8312	15.5
1	$\frac{.2873}{2.560}$	.1122	3	.8329	17.5
5	$\frac{-.2824}{2.5870}$	-.1092	-3	.8343	19.5
11	$\frac{.2238}{2.8516}$	.0785	2	.8355	21.5
3	$\frac{.1208}{2.5596}$	.0472	1	.8358	23.5
7	$\frac{.1066}{8.6005}$	.0124	$\frac{1}{4}$	.8360	25.5

Inspection of the table will show that the correlation increases fairly rapidly with the addition of the first few tests, but after the addition of the sixth test, less than .01 increase is obtained by the addition of six more tests. It was decided therefore to use the first six tests which have a multiple ratio correlation coefficient of .83 as the team to compare with the

other methods. These six tests were accordingly weighted by the weights given in column three of Table XVI. It has been mentioned that these final weights are only approximations to the derived weights. It might be thought that this would affect the correlation obtained when these tests are actually weighted by them. This has sometimes been the case, and in the actual use of this method it has been considered satisfactory if the obtained correlation is within .01 or .02 of that predicted by the formulae. In this case, however, the obtained correlation when these weights were used was also found to be .83. Since this is admittedly only an approximation to a multiple correlation coefficient, it was thought that it would be of interest to find how near the true multiple correlation coefficient it was. These weights were accordingly used as the first

TABLE XVII

RATIO OF DIFFERENCE TO P.E. OF DIFFERENCE FOR DIFFERENT METHODS

Groups	13-year	13-yr. vs.
(Sp., III-V) vs. (VI-IX)	(Sp., III-V) vs. (VIII-IX)	14-yr.
Method		
Raw	18.78	26.13
σ	18.04	24.82
$\sigma \times r_c$	17.26	24.34
$\sigma \times r_I$	18.71	25.45
M.R.	21.76	27.95

estimated weights and Kelley's method of successive approximation⁴ was used to obtain the multiple correlation coefficient. The coefficient was .83 which supports the claim that the multiple ratio correlation coefficient is a close approximation to it.

The results for this team of tests were then compared with those for the other methods. The ratio of the difference between the averages to the P.E. of this difference was found for the groups within the thirteen-year group and for the thirteen- and fourteen-year groups. These ratios are given in Table XVII. The corresponding ratios for the other methods are also included for the sake of comparison. The notation *M.R.* has been used to designate this last method.

In every comparison made, this team of six tests selected and weighted by this method shows more reliable differences than the entire sixteen tests by any of the other methods. It also has a correlation of .83 with the criterion compared with

⁴ *Op. cit.*

a correlation of .80 or .79 for the entire sixteen tests for the other methods. The one respect in which it does not compare favorably is in the matter of reliability.

For the purpose of getting the reliability, two comparable half-scales were built up from the tests used in this team and the other tests included in Table XVI. These half-scales, together with their cumulative correlations, are given in Table XVIII. For example, for the team of six tests the two half-scales consist of tests 10, 13, and 12 with a correlation of .787 and tests 9, 7, and 16 with a correlation of .795.

TABLE XVIII
COMPARABLE HALF-SCALES BUILT UP FROM TESTS IN TABLE XVI

Tests	Half-Scale I	Tests	Half-Scale II
	Cumulative Correlation		Cumulative Correlation
10	.672	9	.699
13	.763	7	.774
12	.787	16	.795
1	.797	4	.804
14	.804	2	.804
3	.806	6	.808

The reliability coefficient obtained for the team of six tests, using these half-scales and applying the Brown-Spearman formula as described in Section III, was .88. It will be remembered that the corresponding coefficients for the sixteen tests were .92 to .93 for the different methods for age thirteen. So that in order to have the results for this team as reliable as in the case of the others, it would be necessary to increase the number of tests or to lengthen the tests already included.

Tests were accordingly added to see what the effect would be on the reliability coefficient. The team was increased to eight, to ten, and to twelve tests, two being added in each case so that one might be added to each half-scale. Table XVI shows how the correlation with the criterion increased with the addition of each test and Table XVIII, which tests composed the half-scales and how these compared. For example, the addition of tests 2 and 1 to the six-test team raised the correlation coefficient from .829 to .833. The half-scales used in obtaining the measure of reliability were the first four in each group in Table XVIII, tests 10, 13, 12, and 1 having a correlation of .797 and 9, 7, 16, and 4 having a correlation of .804. The re-

liability coefficient for this eight-test team is .90. This coefficient would ordinarily be expected to increase with the addition of further tests, but the addition of four more tests did not raise it above this. The reason for this is apparent from an examination of the weights at which these tests were added. The values have decreased to such an extent that the addition of these tests had practically no influence on the total. Another and more usual means of increasing the reliability is to lengthen the tests already included in the team. The formula used above for getting the reliability is often used for the purpose of predicting how much the length of a test should be increased in order to raise a given reliability to any desired reliability. For example, if a reliability of .93 is desired for the eight-test team, the formula before solution would be:

$$.93 = \frac{n (.90)}{1 + (n - 1) .90}$$

The solution of the formula results in the prediction that a reliability of .93 would be obtained if the eight-test team were extended to 1.48 times its present length. The cumulative time for these eight tests is 17.5 minutes. The required length would then be 25.9 minutes. Holzinger⁵ has shown that in one case, at least, the obtained reliability does not increase as rapidly as the predicted reliability after the addition of the first few tests. It might be necessary therefore to extend the test to somewhat more than the predicted length in order to obtain the desired reliability.

Two of the possibilities which this technique shows are: (1) out of the sixteen tests, a team of six tests can be built up requiring 14 minutes work on the part of subjects, exclusive of time for directions, and having a correlation of at least .83⁶ with the criterion and a reliability of .88, or (2) a team of eight tests requiring 17.5 minutes work having a correlation of at least .83 with the criterion and a reliability of .90. This may be compared with the sixteen tests requiring 37.5 minutes work and having a correlation of .80 with the criterion and a reliability of .93. The larger number of tests results in a higher reliability but the correlation with the criterion is

⁵ Holzinger, Karl J., *Note on the Use of Spearman's Prophecy Formula for Reliability*. Jr. Ed. Ps., 14: 302-305, 1923.

⁶ As pointed out above, it is thought that correlation is lower than the true correlation because of errors in grade placement.

lower. Of the two, a higher validity is of course the more important.

It would hardly be expected that many tests of the "pencil and paper" variety could be found without some common elements. Objective examination of the tests is not sufficient to determine this, for two tests having the appearance of being relatively independent may be found to correlate rather highly, and other tests which appear somewhat similar may have low inter-correlations. Adding similar tests to a team of tests may and does have the effect of increasing the reliability, but where the test adds no new elements to the composite, this increase in reliability can be as satisfactorily and more readily obtained by increasing the length of the tests already in the composite.

These comparisons have all been made between the weighted six-test composite and the entire group of sixteen tests scored by the different methods. The superiority of the former is due to two factors, the weighting and the selection of tests. In order to determine the relative importance of these two factors, it would be necessary to compare the weighted six-test composite with the raw score composite for these same six tests. Any superiority then manifested would be conclusive evidence of the superiority of this method of weighting. This comparison was accordingly made. The correlation of the raw score total on these six tests with grade position was found to be .81 compared with .83 for the six-test weighted composite and .79 or .80 for the different methods with the entire group of sixteen tests. The other comparisons between the weighted composite and the raw score total for these tests are included in Table XXIX.

TABLE XXIX

COMPARISON OF WEIGHTED AND RAW-SCORES FOR SIX-TEST COMPOSITE
RATIO OF DIFFERENCE TO P.E. OF DIFFERENCE

Groups	13-yr.		13-yr. vs. 14-yr.
	(Sp., III-V) vs. (VI-IX)	(Sp., III-V) vs. (VIII-IX)	
Weighted Scores	21.76	27.95	8.38
Raw Scores	22.75	29.62	9.24

Inspection of these results shows that the weighted scores have no advantage over the raw scores for distinguishing between these various groups. In each case the raw score total

shows larger and more reliable differences. The advantage gained by the weighted composite is the higher correlation with the criterion, but the general superiority of this six-test composite over the group of sixteen tests scored in the various ways is thus seen to be chiefly due to the tests selected for the composite and less to the weights used for these tests.

The superiority of this technique is then chiefly due to the fact that tests are selected according to the extent to which they increase the correlation of the composite with the criterion and according to their intercorrelations. Any of the methods of weighting considered above are inadequate because they fail to take these correlations into consideration. This would be equally true of any system of weighting which failed to do this. That these methods compared as favorably as they do with the method considered here is due much more to the care used in the selection and arrangement of the tests than to any scoring device which has been used. It cannot be questioned that these tests have stood this close scrutiny extremely well. It is doubtful if some of the tests now on the market which have been more hastily and less discriminately assembled would be able to endure as penetrating an examination. The conclusion that improvements can be made should not obscure the evidences of real worth which have been revealed. The use of this scale can be impartially recommended.

V

SUMMARY

1. Investigation of the three methods of weighting, sigma weighting, sigma $\times r_c$ weighting, and sigma $\times r_I$ weighting with an unselected group of thirteen-year olds failed to reveal any evidence of superiority of these methods over the raw score method for the purpose of obtaining either a higher validity or greater reliability.

2. Comparison of the above group with an unselected group of fourteen-year olds revealed evidence in favor of the methods of weighting. Sigma weighting and sigma $\times r_c$ weighting were apparently slightly more successful in distinguishing between these two groups than the raw score method, while sigma $\times r_I$ weighting was noticeably better in this respect than the raw score method. There was again no evidence of greater reliability for any of the methods.

3. When this study was extended to include the comparison of the two methods of weighting, sigma and sigma $\times r_c$ weighting, with an unselected group of eight-year olds, the results bore out the conclusions made for the comparisons within the thirteen-year group. Exceptions to this were so slight that they may have been due to chance factors not taken into consideration. Even if this were not the case they were not considered large enough to be significant.

4. While there is some discrepancy between the results within unselected age-groups and between unselected age-groups, it is thought that results for the former are more significant. Even if this is not granted, it cannot be concluded that sufficient superiority of any of these methods over the raw score method has been demonstrated to warrant the recommendation of their adoption in preference to the raw score method. This is especially true when the increased expenditure of time required for these methods in both standardization of tests and scoring of tests is taken into consideration.

5. A method of deriving weights by means of multiple correlation was used to build up a team of tests to determine the advantages of this method over the other methods and the raw score method. By this means a six-test composite was built up which required 14 minutes work on the part of the subject, exclusive of time for directions, and which has a correlation of

.83 with the criterion and a reliability of .88. Results for this six-test team also distinguish more successfully than the results for any of the other methods between dull and bright children within the thirteen-year group and between the thirteen-year and fourteen-year groups. This is the first time these comparisons have been consistently in favor of any one method. Another possibility is an eight-test team requiring 17.5 minutes work, having the same correlation with the criterion and a reliability of .90. These may be compared with 37.5 minutes for the sixteen tests, a correlation of .79 or .80 with the criterion, and a reliability of .92 or .93 for the other methods. According to the prediction of the Brown-Spearman formula, the reliability of the eight-test team could be increased to .93 by increasing the length of this test to about 26 minutes.

6. That the advantages of the above method are due chiefly to the selection of the tests which contribute independent elements to the composite, rather than to the derived weights, was demonstrated by means of a comparison of the raw and weighted scores for the six-test composite. The former had a correlation of .81 with the criterion compared with .83 for the latter. But the comparisons of the differences between the various groups for the raw and weighted scores resulted in larger and more reliable differences for the former, thus proving that the advantages of this six-test composite found in the comparisons with the group of sixteen tests are due to the superiority of the tests themselves rather than to the weights used.

7. Probably the chief reason the other methods failed to reveal any superiority over the raw score method is their failure to take into consideration the inter-correlations of tests. Any method of weighting is inadequate which fails to do this. The extent to which tests do or do not contain common elements cannot be determined by merely examining them objectively.

8. Incidentally, attention has been called to the fact that these tests have stood up very well under this examination. Although possibility of improvement has been demonstrated and the direction of improvement indicated, this should not obscure the evidences of real value which have been revealed. It is considered that these have been sufficient to warrant the recommendation of the use of this scale, either instead of others or to supplement others.

BIBLIOGRAPHY

- Buckingham, B. R., "Intelligence and Its Measurement: A Symposium," *Jr. Ed. Ps.* 12: 271-295, 1921.
- Chapman, J. Crosby and Barbara Dale, "A Further Criterion for the Selection of Mental Test Elements," *Jr. Ed. Ps.*, 13: 267-276, 1922.
- Douglas, Harl R. and Peter L. Spencer, "Is It Necessary to Weight Exercises in Standard Tests?" *Jr. Ed. Ps.*, 14: 109-112, 1923.
- Garfiel, Evelyn, *The Measurement of Motor Ability*, Archives of Psychology, 1923.
- Henmon, V. A. C. and Ruth Streitz, "A Comparative Study of Four Intelligence Scales for the Primary Grades," *Jr. Ed. Res.*, 5: 185-194, 1922.
- Henmon, V. A. C., "Intelligence and Its Measurement: A Symposium," *Jr. Ed. Ps.*, 12: 195-198, 1921.
- Holzinger, Karl J., "Note on the Use of Spearman's Prophecy Formula for Reliability," *Jr. Ed. Ps.*, 14: 302-305, 1923.
- Hull, Clark L., "The Joint Yield from Teams of Tests," *Jr. Ed. Ps.*, 14: 396-406, 1923.
- "Intelligence and Its Measurement: A Symposium," *Jr. Ed. Ps.*, 12: 123-154, 195-216, 1921.
- Kelley, Truman Lee, *Statistical Method*, MacMillan, 1923.
- Kelley, Truman Lee, "Report of the Sub-committee on Statistical Methods of the Standardization Committee," *Jr. Ed. Res.*, 4: 77-78, 1921.
- Kelley, Truman Lee, "The Reliability of Test Scores," *Jr. Ed. Res.*, 3: 370-379, 1921.
- Kohs, Samuel C., "Percentile Norms for Scaling Data," *Jr. Ed. Ps.*, 9: 101-102, 1918.
- Kuhlmann, F., *A Handbook of Mental Tests*, Warwick and York, 1922.
- McCall, Wm. A., *How to Measure in Education*, MacMillan, 1922.
- Memoirs*, National Academy of Sciences, Vol. XV, 1921.
- Mohlman, Dora Keene, "The Discriminative Value of the Sub-Tests of a Group Intelligence Scale," *Sch. & Soc.*, 14: 399-400, 1922.
- Morrison, J. Cayce, W. B. Cornell and Ethel Cornell, "A Study of Intelligence Scales for Grades Two and Three," *Jr. Ed. Res.*, 9: 46-56, 1924.
- Otis, Arthur S., "A Method of Inferring the Change in a Correlation Coefficient Resulting from a Change in the Heterogeneity of a Group," *Jr. Ed. Ps.*, 13: 293-4, 1922.
- Rugg, Harold O., *Statistical Methods Applied to Education*, Houghton Mifflin Co., 1917.
- Ruml, Beardsley, "The Need for an Examination of Certain Hypotheses in Mental Tests," *Jr. Phil. Psych. & Sc. Meth.*, 17: 57-61, 1920.
- Thorndike, Edward L., *Mental and Social Measurements*, Teachers College, Columbia University, 1922.
- Thorndike, Edward L., "Fundamental Theorems in Judging Men," *Jr. App. Ps.*, 2: 67-76, 1918.
- Thurstone, L. L., "A Scoring Method for Mental Tests," *Psych. Bull.*, 16: 235-240, 1919.
- West, Paul W., "The Significance of Weighted Scores," *Jr. Ed. Ps.*, 15: 302-308, 1924.
- Will, Harry S., "A Method of Commensurating Mental Measurements," *Jr. Ed. Res.*, 5: 139-153, 1922.
- Woodrow, Herbert and Grace Arthur, "An Absolute Intelligence Scale: A Study in Method," *Jr. App. Ps.*, 3: 118-137, 1919.
- Woody, Clifford, "A Survey in Educational Research in 1923," *Jr. Ed. Res.*, 9: 357-381, 1924.
- Vincent, Leona, *A Study of Intelligence Test Elements*, Teachers College, Columbia University, 1924.

APPENDIX

SAMPLES OF THIRTEEN YEAR TESTS¹

TEST 1²

	A	E	U	B	D	G	C	F	H	
	1	2	3	4	5	6	7	8	9	
Examples:	(A)	1	6	2		(B)	8	1	7	2
	(1)	5	3	6						Time—2 min.
	(2)	9	1	5						
	(3)	5	3	2						
	(4)	4	1	6						
	(5)	2	1	7	9					
	(6)	9	3	6	2					
	(7)	4	2	1	5					
	(8)	1	6	3	2					
	(9)	7	9	1	8	2				
	(10)	8	3	5	6	2				
	(11)	4	2	5	1	3	4			
	(12)	5	2	4	1	3	7	9		

TEST 2

Examples:	free	good	old	heavy	bad	fast	Time 1½ min.
	early	slow	wrong	light	big	right	
1	old	rich	wide	poor	green	full	
2	light	soon	bad	sick	dark	narrow	
3	brown	open	full	dark	sorry	empty	
4	laugh	now	wait	whistle	study	cry	
5	soon	above	when	even	below	back	
6	strong	fight	weak	muscle	jump	work	
7	like	fun	friend	cousin	enemy	skate	
8	never	where	while	still	quickly	always	
9	sharp	narrow	point	steep	dull	study	
10	string	line	straight	turn	old	crooked	

TEST 3

Examples:	chair	book	couch	desk	box	letter	Time—2 min.
	dog	cheese	dish	potato	table	bread	
1	dirt	iron	force	silver	brick	wire	
2	ship	waves	cart	road	wagon	bricks	
3	store	banana	basket	apple	seed	plum	
4	sea	rock	mountain	lake	storm	river	
5	glass	hat	room	ribbon	basket	dress	
6	robin	winter	horse	song	squirrel	fence	
7	rain	wind	sky	steam	heat	water	
8	brass	piano	violin	party	pleasure	flute	
9	submarine	officer	duty	bomb	trench	gun	
10	poetry	physics	physiology	beauty	chemistry	resonance	

¹ An explanatory footnote is added in those cases where it is not obvious what the task is.

² A substitution test, letters for numbers.

TEST 4

Time—2 min.

Examples:	table	box	<u>furniture</u>	red	cloth	wood
	apple	cherry	seed	grow	fruit	leaf
1	silk	red	pretty	dress	fashion	cloth
2	salmon	meat	water	swim	fish	food
3	sheep	flock	animal	meat	woolly	butchered
4	diamond	precious	value	sparkles	jewel	ring
5	hammer	carpenter	nail	tool	useful	iron
6	lettuce	vegetable	green	leaves	healthful	garden
7	man	boy	strong	fights	muscle	person
8	gun	shoot	muzzle	weapon	dangerous	wound
9	carpentry	tools	trade	man	wages	house
10	gold	bright	valuable	mineral	ring	money
11	wagon	vehicle	brake	wood	ride	carriage
12	baseball	practise	diamond	healthful	team	sport
13	bee	wax	bird	honey	insect	stings
14	mustard	burns	spice	powder	strong	flavor
15	honesty	excellence	best	virtue	right	desirable

TEST 5

Time—2 min.

Examples:	bread	meat	eggs	<u>plate</u>	cheese ³
	bush	stone	tree	flower	grass
1	top	rattle	doll	sled	playing
2	book	marbles	pencil	map	slate
3	cup	saucer	plate	spoon	bowl
4	skating	language	arithmetic	spelling	reading
5	apples	peaches	nuts	pears	cherries
6	mother	cousin	brother	aunt	friend
7	town	house	village	hamlet	city
8	sparrow	butterfly	bee	rabbit	eagle
9	you	we	and	I	he
10	free	happy	glad	joyous	pleased
11	automobile	ship	motorcycle	bicycle	aeroplane
12	general	ensign	major	colonel	captain
13	energetic	ambitious	cautious	industrious	zealous
14	amazement	wonder	surprise	astonishment	anger
15	foolhardy	dangerous	reckless	venturesome	rash

TEST 6

Time—2 min.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Examples:	(A)	The	third	letter	of	the	alphabet	is																		
	(B)	The	second	letter	before	the	sixth	letter	is																	
1	The	fifth	letter	of	the	alphabet	is																			1
2	The	second	letter	before	the	last	letter	is																		2
3	The	third	letter	before	M	is																				3
4	The	letter	midway	between	H	and	N	is																		4
5	The	second	letter	after	the	fourth	letter	is																		5
6	The	letter	two	letters	to	the	right	of	the	letter	E	is														6
7	The	first	letter	to	the	left	of	the	tenth	letter	is															7
8	The	letters	of	the	word	<i>the</i>	in	the	order	in	which	they	come	in	the	alphabet	are									8
9	The	letters	of	the	word	<i>boy</i>	in	the	order	in	which	they	come	in	the	alphabet	are									9
10	The	word	you	get	by	putting	the	first	letter	between	the	two	middle	letters	of	the	alphabet	is								10

³ Draw a line under the one which does not belong with the others in the line.

TEST 7

Time—2 min.

Write one number after each one of these words:

If the word contains the letter
a write 1 after it, if not
 write 2 after it.

Examples: handle 1
 white 2

{	length	willing
	wrong	table
	dread	chest
	ladle	beauty
	comrade	basket

If the word contains
a and *i* write 1 after it
a but not *i* write 2 after it,
i but not *a* write 3 after it.

Examples: gracious 1
 desperate 2
 trickle 3

{	camel	carriage
	triangle	invite
	cautious	national
	simple	around
	fearful	united
	example	peaceful
	forgive	animal
	irritate	accurate
	bicycle	previous
	separate	religious

TEST 8⁴

Time—3 min.

Examples: 7...2 = 5
 4...4 = 8
 2...5 = 10
 9...3 = 3

- | | |
|--------------------|----------------------|
| (1) 2...3 = 6 | (7) 18...3...2 = 3 |
| (2) 7...2 = 9 | (8) 4...3...2 = 24 |
| (3) 8...4 = 2 | (9) 25...5...2 = 7 |
| (4) 9...4 = 13 | (10) 11...2...2 = 20 |
| (5) 9...2...3 = 14 | (11) 5...6 = 3...10 |
| (6) 12...1...6 = 5 | (12) 24...6 = 3...1 |

TEST 9

Time—2 min.

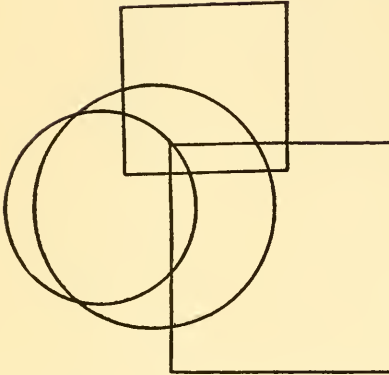
Examples: table top paint legs cloth dishes
 tree shade nuts roots leaves branches

- | | |
|-------------|--|
| 1 book | story pages shelf picture printing |
| 2 squirrel | nuts fur tail cage tree |
| 3 cat | hair owner mouse claws milk |
| 4 chair | arms legs rocker seat comfort |
| 5 house | sidewalk window bed furnace door |
| 6 boy | shoes legs suit head knife |
| 7 room | furniture lamp people walls ceiling |
| 8 concert | encore performer violin singing applause music |
| 9 army | officers tents fighting soldiers ships deaths |
| 10 banquet | music wine guests dancing food laughter |
| 11 fire | alarm flame danger heat fireman insurance |
| 12 blizzard | winds death thunder danger snow wrecks |
| 13 club | banquets meetings committees clubhouse fun members |
| 14 trial | sentence crime defendant judge jury guilt |
| 15 contest | opponents crowds rowing strength rivalry dislike |

⁴ Supply missing signs.

TEST 10

Time—3 min.



Examples:

Write the letter A in the space which is only in the small square.

Write the letter B in the space which is only in the large circle.

1. Write the number 1 in the space which is only in the small circle.
2. Write the number 2 in the space which is only in the large square.
3. Write the number 3 in the space which is only in the small and large squares.
4. Write the number 4 in the space which is only in the small and large circles.
5. Write the number 5 in the space which is only in the small square and large circle.
6. Write the number 6 in the space which is only in the large square and large circle.
7. Write the number 7 in the space which is only in both squares and the large circle.
8. Write the number 8 in the space which is only in both circles and the large square.
9. Draw a line from the upper right corner of the small square to the lower right corner of the large square.
10. Draw a line from the upper left corner of the large square to the center of the small circle.

TEST 11

Time—2 min.

Examples: 2 4 6 8 ~~9~~ 10 12
 9 8 7 3 6 5 4

(1)	3	5	7	8	9	11
(2)	1	4	7	10	12	13
(3)	9	7	4	5	3	1
(4)	18	15	12	9	6	5
(5)	2	5	4	6	8	10
(6)	1	5	9	11	13	17
(7)	12	11	10	8	6	4
(8)	3	6	9	12	14	15
(9)	1	5	10	15	20	25
(10)	2	4	6	8	10	32

TEST 12^a

Time—2 min.

③

□4

~~5~~

31	16	7	6	13	25	17	26
8	46	18	22	10	23	49	32
31	50	29	27	14	35	26	34
44	37	10	38	41	9	43	21
29	19	28	49	25	27	34	6
50	7	26	21	13	29	44	17
23	18	14	8	46	35	11	41
32	31	38	27	47	28	19	35
9	37	25	19	43	31	16	23

^a Draw a circle around the number if 3 goes into it evenly, a square if 4 goes into it evenly, etc.

TEST 13^a

Time—3 min.

Examples: girl come ill his
apple shell ripe banana

1. sit can pie big
2. ton sing boy some
3. tell some me can
4. why bury still you
5. are bat out tell
6. truth happy people riches
7. mirth beauty business ugly
8. trill hurry battle leaves
9. tramp lease trial found
10. across bought camel truce
11. makes story tremble asking
12. early income fashion simply
13. anchor sample truth ripple
14. beacon giving nation humble
15. button forgive angel bought

TEST 14

Time—3 min.

Examples: What is the number which is 2 less than $\frac{1}{3}$ of 9?—

What is the number which if added to 3 is $\frac{1}{2}$ of 12?—

1. What is the number which is 2 more than $\frac{1}{2}$ of 10?.....—
2. What is the number which if multiplied by 2 is 3 times 6?....—
3. What is the number $\frac{1}{3}$ of which is $\frac{1}{5}$ of 15?.....—
4. What is the number which if divided by 2 leaves 1 less than 5?—
5. What is the number which if added to 8 makes 3 less than 15?—
6. What is the number which if multiplied by 2 makes 3 more than 11?—
7. What is the number which if multiplied by itself is $\frac{1}{4}$ of 100?—
8. What is the number $\frac{1}{3}$ of which is $\frac{5}{6}$ of 18?.....—
9. What is the number which if subtracted from 17 leaves 4 more than $\frac{2}{3}$ of 15?.....—
10. What is the number which if added to 9 leaves twice the product of 2 times $\frac{1}{3}$ of 24?.....—

TEST 15

Time—3 min.

Examples: my not is book that
ran the boy the street down

- 1 apples trees on grow
- 2 play boys like marbles to
- 3 grow boys men to become up
- 4 is lesson girl her studying the
- 5 there days are the week in seven
- 6 children room of the out ran six
- 7 away winter for nuts store squirrels
- 8 Mary I runs as as fast
- 9 do go we Saturday school on not to
- 10 she youngest selected our the in girl room
- 11 thousand many a year cars makes Ford
- 12 true stories teacher about the a told them colonies
- 13 who her lost girl pencil the another bought
- 14 allowed upon skate to they never river were the
- 15 an embankment train leaped lost lives their and
many people the

^a Find a letter common to three of the words. Draw a line under the word without that letter.

TEST 16^{*}

Time—3 min.

$$\begin{array}{r} 4 \overline{) 8-6} \\ 224 \end{array}$$

$$\begin{array}{r} 3 \overline{) 792} \\ 2-4 \end{array}$$

$$\begin{array}{r} 8 \overline{) 78-0} \\ 975 \end{array}$$

$$\begin{array}{r} 9 \overline{) 7-05} \\ 845 \end{array}$$

$$\begin{array}{r} 6 \overline{) 8910} \\ 1-85 \end{array}$$

$$\begin{array}{r} 4 \overline{) 9476} \\ 23-9 \end{array}$$

$$\begin{array}{r} 7 \overline{) 5698} \\ 81- \end{array}$$

$$\begin{array}{r} 9 \overline{) -532} \\ 948 \end{array}$$

$$\begin{array}{r} 6 \overline{) 4-74} \\ 829 \end{array}$$

$$\begin{array}{r} 7 \overline{) -376} \\ 768 \end{array}$$

$$\begin{array}{r} 8 \overline{) 6336} \\ 7-2 \end{array}$$

$$\begin{array}{r} 5 \overline{) 8920} \\ 17-4 \end{array}$$

$$\begin{array}{r} 3 \overline{) 59-8} \\ 1996 \end{array}$$

$$\begin{array}{r} 5 \overline{) 9815} \\ 19-3 \end{array}$$

$$\begin{array}{r} 7 \overline{) 9499} \\ 135- \end{array}$$

^{*} Supply missing number.



Classification number

--

Accession number

	4	0	9	3	2
--	---	---	---	---	---

Date Due:

MAY 30 1984

PSYCH

BF21
A7
no.80

Anderson, Rose Gustava, 1893-
A critical examination of test-
scoring methods, by Rose G. Anderson.
New York, 1925.
59 p. 1 illus. 25 cm. (Archives of
psychology ... no. 80)

40932

PSYCH

ARCHIVES OF PSYCHOLOGY

Columbia Univ. P. O., New York City

In addition to the numbers of the Archives, the following monographs are to be obtained from us:

- Measurements of Twins: EDWARD L. THORNDIKE. 50 cents.
Avenarius and the Standpoint of Pure Experience: WENDELL T. BUSH. 75 cents.
The Psychology of Association: FELIX ARNOLD. 50 cents.
The Measurement of Variable Quantities: FRANZ BOAS. 50 cents.
Linguistic Lapses: FREDERIC LYMAN WELLS. \$1.00
The Diurnal Course of Efficiency: HOWARD D. MARSH. 90 cents.
The Time of Perception as a Measure of Differences in Sensations:
VIVIAN ALLEN CHARLES HENMON. 60 cents.
Interests in Relation to Intelligence: LOUISE E. POULL. \$1.00. Reprinted from
Ungraded.
The Conditioned Pupillary and Eyelid Reactions: HULSEY CASON. \$1.00 Re-
printed from the *Journal of Experimental Psychology*.

THE JOURNAL OF PHILOSOPHY

Sub-Station 84, New York City

Published on alternate Thursdays

\$4 PER ANNUM, 26 NUMBERS 20 CENTS PER COPY

Edited by Professors F. J. E. WOODBRIDGE and WENDELL T.
BUSH of Columbia University.

There is no similar journal in the field of scientific philosophy. It is identified with no philosophical tradition, and stands preeminently for the correlation of philosophy with the problems and experience of the present.

Among the many psychological articles appearing in this Journal are the following of recent date:

- Measures of Intelligence: A. T. POFFENBERGER
The Identity of Instinct and Habit: KNIGHT DUNLAP
Must We Give Up Instincts in Psychology? J. R. GEIGER
The Modification of Instinct: WALTER S. HUNTER
Mr. Russell's Psychology: F. C. S. SCHILLER
Intelligence and Intellect: A. A. ROBACK

ARCHIVES OF PHILOSOPHY

Editorial communications should be addressed to
Professor FREDERICK J. E. WOODBRIDGE, Columbia University,
New York City

Order should be sent to ARCHIVES OF PHILOSOPHY
Sub-Station 84, New York City

The numbers are as follows:

- The Concept of Control. SAVILLA ALICE ELKUS. 40 cents.
The Will to Believe as a Basis for the Defense of Religious Faith.
ETTIE STETTMEIER. \$1.
The Individual: A Metaphysical Inquiry. WILLIAM FORBES COOLEY. \$1.00.
The Ethical Implications of Bergson's Philosophy. UNA BARNARD SAIT. \$1.25.
Religious Values and Intellectual Consistency. EDWARD H. REISNER. 75 cents.
Rosmini's Contribution to Ethical Philosophy. JOHN FAYATA BRUNO. 75 cents.
The Ethics of Euripides. RHYS CARPENTER. 50 cents.
The Logic of Bergson. GEORGE WILLIAM PECKHAM. 75 cents.
The Metaphysics of the Supernatural as illustrated by Descartes.
LINA KAHN. \$1.00.
Idea and Essence in the Philosophies of Hobbes and Spinoza.
ALBERT G. A. BALZ. \$1.25.
The Moral and Political Philosophy of John Locke.
STERLING TOWER LAMPRECHT. \$1.50.
Science and Social Progress. HERBERT WALLACE SCHNEIDER. \$1.25.
The Nature of Life: A Study in Metaphysical Analysis.
FLORENCE WEBSTER. \$1.25.
Empedocles' Psychological Doctrine. WALTER VEAZIE. \$1.00.

DIRECTORY OF AMERICAN PSYCHOLOGICAL PERIODICALS

- American Journal of Psychology.** Ithaca, N. Y.: Morrill Hall.
 Subscription \$6.50. 600 pages annually. Edited by E. B. Titchener.
 Quarterly. General and experimental psychology. Founded 1887.
- Pedagogical Seminary and Journal of Genetic Psychology**—Worcester, Mass. Clark University.
 Subscription \$7. 700 pages annually. Edited by Carl Murchinson and an international cooperating board.
 Quarterly. Child behavior, differential and genetic psychology. Founded 1891.
- Psychological Review.** Princeton, N. J.: Psychological Review Company.
 Subscription \$4.25. 480 pages annually.
 Bi-monthly. General. Founded 1894. Edited by Howard C. Warren.
- Psychological Bulletin.** Princeton, N. J.: Psychological Review Company.
 Subscription \$5. 720 pages annually. Psychological literature.
 Monthly. Founded 1904. Edited by Samuel W. Fernberger.
- Psychological Monographs.** Princeton, N. J.: Psychological Review Company.
 Subscription \$5.50 per vol. 500 pp. Founded 1895. Ed. by Shepherd I. Franz.
 Published without fixed dates, each issue one or more researches.
- Psychological Index.** Princeton, N. J.: Psychological Review Company.
 Subscription \$1.50. 200 pp. Founded 1895. Edited by Madison Bentley.
 An annual bibliography of psychological literature.
- Journal of Philosophy.** New York: 515 W. 116 St.
 Subscription \$4. 728 pages per volume. Founded 1904.
 Bi-weekly. Edited by F. J. E. Woodbridge and Wendell T. Bush.
- Archives of Psychology.** Columbia University P. O., New York. Archives of Psychology.
 Subscription \$5. 500 pp. annually. Founded 1906. Edited by R. S. Woodworth.
 Published without fixed dates, each number a single experimental study.
- Journal of Abnormal Psychology and Social Psychology.** Albany, N. Y.
 Subscription \$5. Boyd Printing & Publishing Co. Edited by Morton Prince.
 Bi-monthly. 432 pages annually. Founded 1906. Abnormal and social.
- Psychological Clinic.** Philadelphia: Psychological Clinic Press.
 Subscription \$2.50. 288 pages. Ed. by Lightner Witmer. Founded 1907.
 Without fixed dates (9 numbers). Orthogenics, psychology, hygiene.
- Training School Bulletin.** Vineland, N. J.; The Training School.
 Subscription \$1. 160 pages annually. Edited by E. R. Johnstone. Founded 1904.
 Monthly (10 numbers). Psychology and training of defectives.
- Journal of Educational Psychology.** Baltimore: Warwick & York.
 Subscription \$4. 576 pages annually. Founded 1910.
 Monthly (9 numbers). Managing Editor, Harold O. Rugg.
 (Educational Psychology Monographs.
 Published separately at varying prices. Same publishers.)
- Comparative Psychology Monographs.** Baltimore: Williams & Wilkins Co.
 Subscription \$5. 500 pages per volume. Edited by W. S. Hunter.
 Published without fixed dates, each number a single research.
- Psychoanalytic Review.** Washington, D. C.: 3617 10th Street, N. W.
 Subscription \$6. 500 pages annually. Psychoanalysis.
 Quarterly. Founded 1913. Edited by W. A. White and S. E. Jelliffe.
- Journal of Experimental Psychology.** Princeton, N. J.
 Psychological Review Company. 480 pages annually. Experimental.
 Subscription \$4.25. Founded 1916. Bi-monthly. Ed. by John B. Watson.
- Journal of Applied Psychology.** Bloomington: Indiana University Press.
 Subscription \$5.00. 400 pages annually. Founded 1917.
 Quarterly. Edited by James P. Porter and William F. Book.
- Journal of Comparative Psychology.** Baltimore: Williams & Wilkins Co.
 Subscription \$5. 500 pages per volume. Founded 1921.
 Bi-monthly. Edited by Knight Dunlap and Robert M. Yerkes.

BF 21.A7 no80



3 9358 00040932 3

This book may be kept.....*one*.....weeks.

A fine of two cents will be charged for each day books or magazines are kept overtime.

Two books may be borrowed from the Library at one time.

Any book injured or lost shall be paid for by the person to whom it is charged.

No member shall transfer his right to use the Library to any other person.

Classification number

BF 21 A7 no.80

40932

Anderson, Rose Gustava
A critical examination of
test- scoring methods.

Date	Issued to
MAY 30 1957	C. FOLDS 026 46 99.7
	EXC. LIBRARY ST #9

PSYCH

Anderson, Rose Gustava

BF

A critical examination of test-
scoring methods.

21

A7

40932

No. 80

MBNU

BF 21.A7 no80



3 9358 00040932 3